International Journal of Computer Applications Technology and Research
Volume 5– Issue 7, 417 - 421, 2016, ISSN:- 2319–8656

Development of Computational Tool for Lung Cancer Prediction Using Data Mining

Divya Chauhan
Shoolini University
Solan, Himachal Pradesh
India

Varun Jaiswal
Shoolini University
Solan, Himachal Pradesh
India

Abstract: The requirement for computerization of detection of lung cancer disease arises ever since recent-techniques which involve manual-examination of the blood smear as the first step toward diagnosis. This is quite time-consuming, and their accurateness depends upon the ability of operator's. So, prevention of lung cancer is very essential. This paper has surveyed various techniques used by previous authors like ANN (Artificial Neural Network), image processing, LDA (Linear Dependent Analysis), SOM (Self Organizing Map) etc.

Keywords: Lung Cancer, Classification, Neural Network, SOM, LDA, PCA, Chi-Square, Feature Extraction.

1. INTRODUCTION

1.1 Background

Lung cancer research is one of the most concerning area of interest in medical field. The early diagnose of the cancer can help in increasing the mortality rate of humans [1]. Lung cancer is customarily a contagion which takes place because of the element linked with unimpeded cell or conveniently progress in zones present in lung area. According to American Cancer Society it is approximated that 48,610 persons (27,880 men and 20,730 women) will be detected with and 23,720 men and women will have high percentage of lung cancer in 2013 only [2]. In turn, it is part of the even broader set of diseases disturbing the tuberculosis, Silicosis and Interstitial Lung Disease (ILD), which are all known as diffuse parenchymal lung disease (DPLD) [3].

1.2 Data Mining in Medical Field

Data mining is the process in which valuable information is extracted from the large dataset. It has reached the high growth over past few years. Due to the usefulness of data mining approaches in health world, it has become the good technology in healthcare domain [4]. This realization leads to explosion of data mining approaches [5]. Medical data mining can exploit the hidden patterns present in voluminous medical data which otherwise is left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain [6]. Nevertheless, as on today lot of research is found in the literature. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of diseases, predicting the severity of various diseases, and remote health monitoring. Especially the data mining techniques are more useful in predicting heart diseases, lung cancer, and breast cancer and so on.

1.3 CET Images and its Importance in Medical Field

A CET scanner uses the digital processing to get 3-D image of an object [7]. A CET scanner emits the radiation from a device then scans the whole body to get 3-D image [8]. CET scan is very important as CET scans are a valuable diagnostic tool. They are able to detect some conditions that conventional XY-rays cannot, since CET scans can show a "3-D" view of the section of the body being studied. CET scans are also useful for monitoring a patient's progress during or after treatment [9].

In this paper, various techniques to detect lung cancer will be presented along with brief outline of lung cancer detection.

2. RELATED WORK

Hossein GhayoumiZadeh, et al. [10], 2013 represented an image analysis approach for automated detection, preprocessing-smoothing, enhancement, segmentation, feature extraction-morphological and calorimetric and then detection and categorization of particular cells, particularly the cancer cells from usual cells is complete.

Lim Huey Nee, et al. [11], 2012 presented the incline scale, thresholding, morphological operation and division change to perform cell segmentation. In this paper 50 imageries were used to test the planned method and the effect showed that the process hasmanaged to obtain qualitatively good segmentation consequences.

FauziahKasmin [12], 2012 presented the recognition of blood disorder is through visual inspection of tiny images of blood cells. From the recognition of blood disorders, it can lead to classification of certain diseases related to blood. This document describes a first round study of developing a detection of leukemia types using microscopic blood sample imagery. Here, analyzing through images is very significant as from images; disease can be detected and diagnosed at earlier stage. From there, further actions like scheming, monitoring and prevention of diseases can be done. Imagery is used as they are despicable and do not need expensive testing and lab equipment's. The system will focus on white blood cells disease, leukemia. The system will use features in microscopic images and look at changes on texture, geometry, color and statistical analysis. Changes in these features will be used as a classifier input. A text appraisal has been done and Reinforcement Learning is proposed to classify types of leukemia. A small conversation about issues concerned by researchers also has been ready.

WaidahIsmail [13], 2011 presented a method for the detection and classification of blast cells in M3 with others sub types using computer generated annealing and neural networks. In this paper, we greater than before our test result from 10 images to 20 images. We perform Hill Climbing; Simulated Annealing and Genetic Algorithms for detect the blast cells. As a result,

simulated annealing is the "best" heuristic search for detecting the leukemia cells. From the detection, we perform features extraction on the blast cells and we classify based on M3 and other sub-types using neural networks. We received persuasive result which has targeting around 97% in classify of M3 with other sub-types. Our consequences are based on real world image data from a Hematology Department

3. VARIOUS TECHNIQUES FOR CANCER DETECTION AND PREVENTION

3.1 ANN (Artificial Neural Network)

An artificial neural network does not shot to be like the thought process and if/ then sense of the people brain as completed by an expert system. It mimics exact aspects of the in turn dispensation and objective sympathetic of the brain by means of a network of neural link [14]. As a result, a number of writers record it as a "microscopic", "white box" structure and a professional system as a "macroscopic", "black box" system. An Artificial Neural Network consists of a huge amount of simple dispensation elements that are dependable and covered [15].

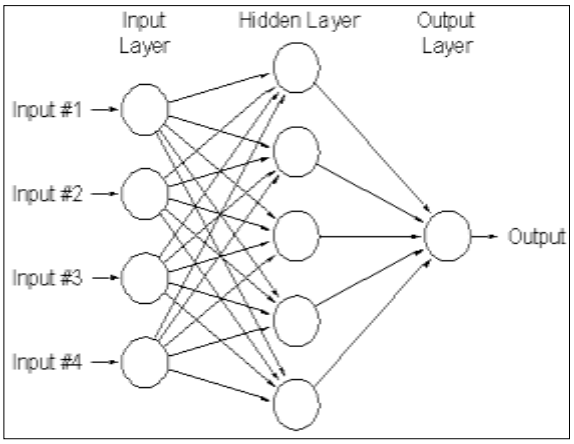


Figure 1: Basic Diagram of A.N.N

Inputs: x1, x2, x3, x4…….xn

Weights: w1j, w2j, w3j, w4j……..wnj

TransferFunction: ∑

Activation Function: α

Output: x1w1j, x2w2jj…….xnwnj

3.2 LDA (Linear Dependent Analysis)

Linear Discriminant Analysis is utmost commonly utilized as dimensionality lessening method in the pre-processing stage for machine learning applications in addition to design-classification. The main objective is to project a specific dataset on top of a lower-dimensional space using virtuous class reparability so as to decrease computational prices as well as also evade overfitting [16]. The novel linear discriminant was first designated for a two-class issue, in addition it was then afterwards widespread as "Multiple Discriminant Analysis" or "multi-class LDA" through C. R. Rao in the year of 1948. Linear Discriminant Analysis is "controlled" as well as calculates the guidelines ("linear discriminants") which would probably signify the axes that are applied to make the most of the separation amongst multiple type of classes. Below are the five basic steps utilized for implementing a LDA technique [17].

A necessary and sufficient condition for the set of functions:

$f_1(x)$, $f_2(x)$...$fn(x)$ to be linearly independent is that

$c_1 f_1(x) + c_2 f_2(x) + ... + c_n f_n(x) = 0$

only when all the scalars c_i are zero.

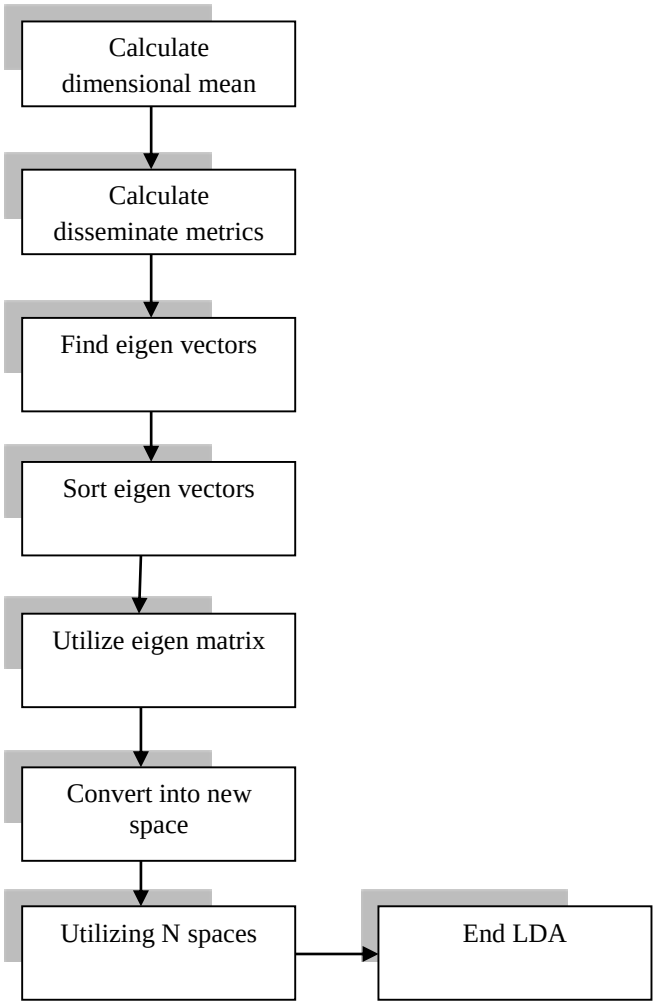


Figure 2: Basic Diagram of L.D.A

3.3 Self-Organizing Map (SOM)

The Self-Organizing Map is one of the commonly used network model. It belongs to the learning networks. The Self-Organizing Map is un-supervised learning method. If Self-Organizing Map is used for feature extraction then it is called Self-Organizing Feature Map [18].

Below figure shows that there are 5 cluster units, Yi and 7 input units, Xi. Clusters are arranged in linear array [19].

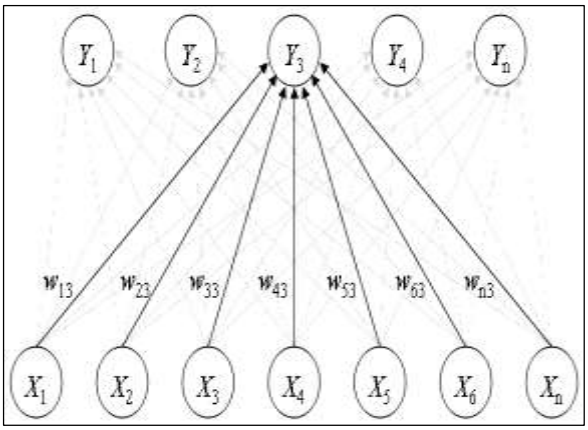


Figure 3: SOM Example

Self-Organizing Map was designed by Kohonen. The SOM has been useful in many applications. It maps the high dimensional space to map units for preserve mapping. Neuron units commonly made lattice onto a plane. Preserving property means reserving the distance between points. In addition to that Self-Organizing Map has the capability of generalizing. It means recognizing the patterns that never met before. The Self-Organizing Map I 2-D can be represented as following**:**

$$Y = \{ x \ldots\ldots x_{acw} \} \qquad (1)$$

The neurons are connected to adjacent neurons by a relation. Commonly, the neurons are connected to each other via rectangular or hexagonal topology. Topologies of neurons are represented above.

Randomly choose a vector

Determine output node wi.

wi x >= wk

Weight update is given as below:

w(new) = w(old) + υ

3.4 Support Vector Machines (SVM)

Support Vector Machine (SVM) is first and foremost a classier technique which executes classification tasks through building hyperplanes in a multi-dimensional space, which divides cases of different and dissimilar class labels. We can define the matrix

$(H)_{ij} = y_i y_j (x_i \cdot x_j), \qquad (2)$

And introduce more compact notation [20]:

Minimize:

$W(a) = -a^T 1 + \frac{1}{2} a^T H a$

Subject to:

$a^T y = 0$

$0 \le a \le C1$

Support Vector machines are also called kernel machines and they have two phases of training:

- Transform input data to high dimensional data.
- Solve quadratic problem [21].

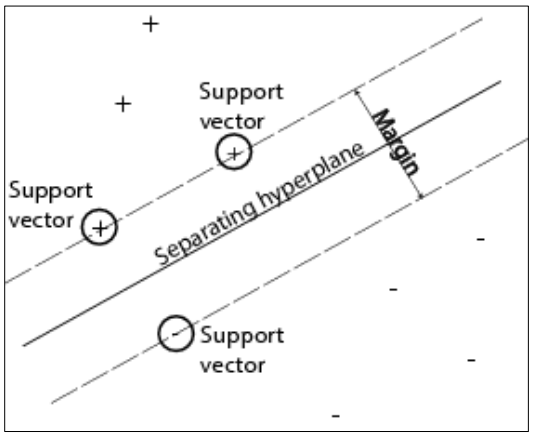


Figure 4: SVM Planar Division

3.5 Genetic Algorithm (GA)

Genetic algorithm is the type of algorithm that is used to solve both constrained and non-constraint problems based on selection criteria. Genetic algorithm modifies the new population and generate new solutions until best solution has not been reached. From large set of population, genetic algorithm uses the random chromosomes to make it parent then make it to produce children [22].

Choose initial population

From left population, select individual chromosomes.

Choose best selected chromosomes

Do crossover

Do repetition

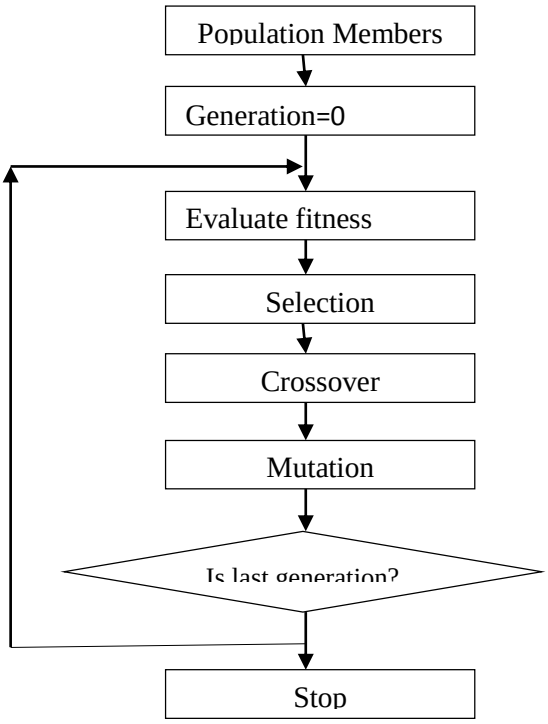
End

Figure 5: Genetic Algorithm Process

3.6 Principal Component Analysis (PCA)

Principal components analysis (PCA) is basically useful for reducing the number of variables that consists a data set while

retaining the inconsistency in the data and to identify unknown patterns in the data and to classify them according to how much of the information, stored in the data, they report for [23].



PCA allows calculating a linear alteration that maps information as of a high dimensional space to a lower dimensional space [24].

B1 = t11 a11 +……t1n an

B2 = t21 a1 +……..t2an

Linear transformation implied by PCA

The linear transformation R^N -> R^K3 that performs the dimensionally reduction

B1 U^t1

B2 U^t2 (x- |x|) = U^t (x- |x|)

3.7 Discreet Wavelet Transform (DWT)

With the enlargement in utilization of internet, communication of data has turn out to be quiet easy. In contrast with the data communication in analog form, digital communication offers us several aids for instance enhanced/superior quality, high speed, compression of data etc [25]. However, image acquisition has some shortcomings also, such as the noise present during transmission. The recognition of the specific data is one of the significant necessities in the arena of information transmission, whether it is the transmission of information/data in military-applications or transmission of pictures on internet that desires to be safer than before [26].

The wavelet transform has grown pervasively approval in denoising of image as well as signal processing. It is the breaking down a specific signal into scaled along with shifted versions of the unique wavelet. A wavelet is a type of waveform of efficiently restricted duration which has average value of zero. And for signals; the identity of the specific signal is specified through the component of low-frequency.

We can approximate a discrete signal in k^2 (X) [1] by

$$f[b] = \frac{1}{\sqrt{N}}\sum_j \; Q_\phi[h_0, j]\, \phi_{h_0,j}[b] + \frac{1}{\sqrt{N}}\sum_{h=h_0}^{\infty}\sum_j \; Q_\psi[h, j]\psi_{h,j}[b] \quad (3)$$

Here, $f[b], \phi_{h_0,j}[b]\; and\; \psi_{h,k}[b]$ are discrete functions which are defined in [0, N-1], to-tally N points. For the reason that the sets $\{\phi_{h_0,j}[b]\}_{j\in X}$ and $\{\psi_{h,j}[b]\}_{(h,j)\in X^2, h \geq j_0}$ are orthogonal to each other. We can simply take the inner product to obtain the wavelet coefficients:

$$Q_\phi[h_0, j] = \frac{1}{\sqrt{N}}\sum_b \; f[b]\, \phi_{h_0,j}[b] \quad (4)$$

$$Q_\psi[h_0, j] = \frac{1}{\sqrt{N}}\sum_b \; f[b]\, \psi_{h,j}[b] \quad h \geq h_0 \quad (5)$$

(4) are called approximation coefficients while (5) are called detailed coefficients.

3.8 Chi Square Test Analysis

The chi-squared one-variable test serve a principle comparable to the binomial test, excluding that it can be used when there are more than two categories to the variable. Thus, if you want to resolve if the numbers of people in each of several categories vary from some predict values, the chi-squared one-variable test is proper. The chi-square goodness-of-fit test is a single-sample non-parametric test, also referred to as the one-sample goodness-of-fit test [27].

4. CONCLUSION AND FUTURE SCOPE

Lung cancer is one of the major health problems in all over world. Cancer constitutes 10.3% of medically certified deaths, which is the most leading cause of death after disease of the circulatory system, accidents and disease of the respiratory system. There are over 100 different types of cancer and one of them is lung cancer. In lung cancer treatment delay results in high mortality rate. So, this paper has reviewed cancer cell detection using various methods.

Use of support vector machines will be considered in the future work as a classification tool. Support Vector Machine (SVM) is also called Support Vector Networks are supervised learning models that analyze data and recognize patterns.

5. REFERENCES

[1] Raje, C.; Rangole, J., "Detection of Leukemia in microscopic images using image processing," in Communications and Signal Processing (ICCSP), 2014 International Conference on , vol., no., pp.255-259, 3-5 April 2014.

[2] Kalyanmoy Deb, A. Raji Reddy, "Reliable classification of two-class cancer datausing evolutionary algorithms", Elsevier, BioSystems , Vol.72, pp.111–129, 2003.

[3] SubrajeetMohapatra, SushantaShekharSamanta, DiptiPatra and SanghamitraSatpathi, "Fuzzy based Blood Image Segmentation for Automated Leukemia Detection", IEEE, 2012.

[4] Nimesh Patel, AshotoshMehra, "Automated Detection of Leukimiausing microscopic images", Elsevier, Vo. 58, 2015.

[5] Jafar, I., Hao Ying , Shields ,A.F., Muzik , O. 'Computerized Detection of Lung Tumors in PET/CT Images', EMBS 2006, 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006.

[6] Nakao M, Kawashima A, Kokubo M, Minato K. "Simulating Lung Tumor Motion for Dynamic Tumor-Tracking Irradiation". Nuclear Science Symposium Conference Record, 2007. NSS 2007

[7] E, Donald, "Introduction to Data Mining for Medical Informatics," Clin Lab Med, pp. 9-35, 2008.

[8] R. Zhang, Y, Katta, "Medical Data Mining," Data Miningand Knowledge Discovery, pp. 305-308, 2002.

[9] Irene M. Mullins et al., "Data mining and clinical data repositories: Insights from a667,000 patient data set," Computers in Biology and Medicine, vol. 36, pp. 1351-1377, 2006.

[10] Zadeh, Hossein Ghayoumi, SiamakJanianpour, and JavadHaddadnia, "Recognition and Classification of the Cancer Cells by Using Image Processing and Lab VIEW," International Journal of Computer Theory and Engineering (2013).

[11] L. H. Nee, M. Y. Mashor, R. Hassan,"White Blood Cell Segmentation for Acute Leukemia Bone Marrow Images," International Conference on Biomedical Engineering (ICoBE),Penang, Malaysia, 27-28 February 2012.

[12] Kasmin, Fauziah, Anton SatriaPrabuwono, and Azizi Abdullah, "Detection ofLeukemia in Human Blood Sample Based On Microscopic Images: A Study, " Journal of Theoretical & Applied Information Technology46.2 (2012).

[13] Ismail, Waidah, et al. "The detection and classification of blast cell in Leukaemia Acute PromyelocyticLeukaemia (AML M3) blood using simulated annealing and neuralnetworks." (2011).

[14] K.A.G. Udeshani, R.G.N. Meegama, T.G.I. Fernando, "Statistical Feature-based Neural Network Approach for the Detection of Lung Cancer in Chest X-Ray Images," International Journal of Image Processing (IJIP), Volume (5), Issue (4) , 2011.

[15] Jinsa , "Lung cancer classification using neural networks for CT images", Computer Methods and Programs in Biomedicine, Volume 113, Issue 1, January 2014, Pages 202-209

[16] J. Yang, D. Zhang, J.-Y. Yang and B. Niu , "Globally max-imizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics" , IEEE transactions on pattern analysis and machine intelligence , vol. 29 , no. 4 , pp.650 -664 , 2007

[17] Young Tae Lee; Yong Joon Shin; Cheong Hee Park, "Extending Linear Discriminant Analysis by Using Unlabeled Data," in Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on , vol., no., pp.557-562, Aug. 31 2011-Sept. 2 2011.

[18] C. J. Lin, C.H. Chu, C.Y. Lee, Y.T. Huang, "2D/3D Face Recognition Using Neural Networks Based on Hybrid Taguchi Particle Swarm Optimization", Eighth International Conference on Intelligent Systems Design and Application (ISDA), 307-312, DOI : 10.1109/ISDA.2008.286.

[19] Timothy Rumbell, Susan L. Denham, and Thomas Wennekers, "A Spiking Self-Organizing Map Combining STDP, Oscillations, and Continuous Learning", IEEE Transactions On Neural Networks And Learning Systems, Vol. 25, No. 5, May 2014

[20] M. Hearst. Support vector machines. IEEE Transactions on IntelligentSystems, 18 – 28, 1998.

[21] Detection of Lung Nodule Using Multiscale Wavelets and Support Vector Machine. K.P.Aarthy, U.S.Ragupathy

[22] Man, K.F.; Tang, K.S.; Kwong, S., "Genetic algorithms: concepts and applications [in engineering design]," in Industrial Electronics, IEEE Transactions on , vol.43, no.5, pp.519-534, Oct 1996, doi: 10.1109/41.538609.

[23] Taranpreet Singh Ruprah, "Face Recognition Based on PCA Algorithm," Special Issue of International Journal of Computer Science & Informatics (IJCSI), 2231–5292, Vol.- II, Issue-1, 2

[24] M. Turk and A. Pentland. Eigenfaces for face recognition, "Cognitive Neuroscience Journal," vol. 3, no. 1, pp.71-86, March 1991.

[25] Z. Tufekci and J. N. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," IEEE Inter.Conf. Southeastcon2000, pp. 116-123, April 2000.

[26] K. P. Soman and K. I. Ramchandran, Insight into Wavelets from Theory to Practice, Printice-Hall of India, 2e, 2005.

[27] Yong Li, "Applications of Chi-Square Test and Contingency Table Analysis in Customer Satisfaction and Empirical Analyses," in Innovation Management, 2009. ICIM '09. International Conference on , vol., no., pp.105-107, 8-9 Dec. 2009